ChinaXiv合作期刊 图书情报工作
第64卷 第21期 2020年11月

异质信息网络嵌入视角下公安微博传播预测研究*

孙冉[1] 安璐[1,2]

[1] 武汉大学信息管理学院 武汉 430072 [2] 武汉大学信息资源研究中心 武汉 430072

摘 要：[目的/意义]预测用户是否转发、评论通缉微博，研究及评估影响通缉微博传播的重要特征，有助于公安微博提升其运营绩效，增强警民之间的沟通和合作。[方法/过程]针对通缉微博的特点，在抽取通缉微博的用户特征、时间特征、微博文本结构特征的基础上，提取通缉微博中的案件特征，包含案件地点关键字、时间关键字、通缉令等级、有无悬赏等，利用 xgboost 算法计算不同特征在转发、评论预测中的重要性，并结合传播网络特征和节点属性，构建基于特征属性异质信息网络嵌入的公安微博传播预测模型，并对模型进行训练和评估。[结果/结论]预测模型在转发、评论数据集上的 AUC 值分别达到 0.737 和 0.799。由于该模型融合了网络结构特征和不同节点属性，更贴近现实的异质信息网络，相比传统的链接预测模型精确度更高。另外，特征重要性实验结果表明，所提出的案件关键字特征在影响微博转发、评论预测的所有特征中重要性最高。

关键词：信息传播 公安微博 链接预测 图表示学习 异质信息网络
分类号：G203
DOI：10.13266/j.issn.0252-3116.2020.21.010

1 引言

随着大数据时代的到来，信息传播渠道日益多元化，新浪微博、微信等社交媒体平台已成为国内重要的政务媒体新平台。根据中国互联网络信息中心（CNNIC）发布的《第 44 次中国互联网络发展状况统计报告》显示，截至 2019 年 6 月，我国在线政务服务用户规模达到 5.09 亿，占网民整体的 59.6%，我国已有 297 个地级行政区政府开通了“两微一端”等新媒体传播渠道，总体覆盖率达 88.9%。2019 年 6 月 24 日，一段女子在街头遭男子暴打的视频在网络上引发热传，随后各地警方纷纷介入调查，由于无法找到夜间视频的来源，中国警方在线在微博上向大众征集相关线索，25 日陆续有网友向警方提供线索，晚上 22 时警方将犯罪嫌疑人抓获。这充分体现了网络社区中群众性参与行为获取信息的强大功能，这种网民互助合作的形式为警方收集侦察线索、查破案件提供了新思路。如何在虚拟网络社区开展群众工作，发挥广大群众的作用帮助公安机关惩恶扬善，共同打击犯罪，是我国公安侦查工作需要考虑的问题。

网络空间的信息传播从用户视角来看是个体之间多样化的交互扩散过程，在网络舆情信息传播的研究中，学者多考虑用户之间的转发关系[1]，对于评论则多从文本分析、情感分析等视角进行研究，但用户之间的交互还涉及到关注、评论等，仅仅考虑转发关系并不能准确捕捉用户在舆情事件传播过程中的交互网络。本文以通缉微博为研究对象，采用 xgboost 算法探究微博内容特征、微博文本结构特征等不同特征在微博转发、评论预测问题上的表现；以用户和微博为节点，根据用户和微博、用户和用户之间的关系构建异质信息网络，并结合传播网络结构和节点属性特征，构建基于特定属性的异质信息网络嵌入模型来对公安微博的转发、评论、关注等传播活动进行链接预测，从而提高公安微博传播预测的准确率。研究用户的各种信息传播行为有助于相关部门理解信息传播机制，对于政府部门进行舆情监控、微博个性化推荐等具有重要意义；研究旨

* 本文系教育部哲学社会科学研究重大课题攻关项目“提高反恐怖主义情报信息工作能力对策研究”（项目编号：17JZD034）、国家自然科学基金重大课题“国家安全大数据综合信息集成与分析方法”（项目编号：71790612）和国家自然科学基金创新研究群体项目“信息资源管理”（项目编号：71921002）研究成果之一。

作者简介：孙冉（ORCID：0000-0002-3597-0096），博士研究生；安璐（ORCID：0000-0002-5408-7135），教授，博士生导师，通讯作者，E-mail：anlu97@163.com。

收稿日期：2020-04-27 修回日期：2020-08-04 本文起止页码：67-76 本文责任编辑：易飞

在揭示公安微博信息传播的影响因素和作用机制以及公众的转发、评论行为模式,为公安微博的运营和建设提供建议,有利于政府部门提高公安微博的传播影响力,增强警民联系沟通与合作互动。

2 相关研究

2.1 特定领域的微博传播模式研究

目前,学者们多采用内容分析法[2]、社交网络分析法[3]、神经网络[4]等方法对政务、灾害、健康等特定领域的微博传播规律进行探讨。在宏观视角下,基于转发层级、转发次数等可将政务微博信息传播模式划分为两级传播模式、普通多级传播模式、卫星传播模式等[5];在微观个体视角下,学者们对微博用户交互模式的研究较多,不同主题下灾害信息传播网络的用户交互模式具有一定的差异性[6]。

通缉是公安机关对于应当逮捕的在逃犯罪嫌疑人、被告人或罪犯进行通令缉捕归案的一项侦查措施,它常以发布通缉令的方式进行。近年来,公安微博为公安机关发布信息提供了新的平台,这使得网络通缉令的传播范围更广泛、传播速度更快,从而弥补了传统通缉令的地域限制,也能更快捷地从群众中获取通缉嫌疑犯的有关线索。通过对相关文献进行梳理发现,虽然关于微博传播规律的研究成果较为丰富,但是以通缉信息作为对象,探讨公安微博信息传播过程中不同用户主体微博传播模式特征和传播用户属性的文献较少,已有的研究多数从公安微博的运营管理视角进行归纳总结式的描述性分析[7],而缺乏基于微博用户行为等客观数据的实证研究。

2.2 微博传播预测研究

微博传播预测研究一般可从微博和用户的视角进行:①微博信息的流行度预测,包括微博转发规模、速度、效率等预测,其分析方法一般是基于传染病模型的数学建模方法以及基于机器学习的分类及回归模型方法[8]。徐月梅等[9]基于用户特征、时间特征及内容特征,采用卷积神经网络和梯度提升决策树算法对政务微博的转发规模进行预测,并且找到影响该规模的重要特征;②微博用户的传播行为预测,用户通过不同的方式参与公共生活,如发布对社会问题的意见、评论或者转发他人的内容等[10]。其中,用户的转发行为是微博信息扩散的最重要方式[11-12]。学者多以用户特征和微博文本的主题、结构特征为基础,运用SVM、逻辑回归模型等机器学习算法实现微博用户转发预测[13],J. Zhu等[14]探讨了文本内容、影响力和时间对转发的影响,并利用逻辑回归分类器进行预测,发现加入时间因素更有利于理解转发机制;近年来学者多从用户社会网络关系出发,结合用户行为日志数据,将复杂网络方法和内容分析结合进行用户转发行为预测,B. Liang等[15]基于影响微博转发的基本因素,利用单类协同过滤方法测量用户偏好和影响力来对用户的转发行为进行预测。

学者们主要从微博信息传播影响因素和用户行为影响因素进行指标上的探索,多关注微博信息传播过程中的用户转发行为,对微博评论则多从主题抽取和情感分析等视角进行研究。不同影响因素会导致用户参与行为的不同[16],但目前与之相关研究较少,本文拟综合考虑信息传播过程中用户的转发、评论行为,探讨哪些影响因素会对用户不同参与行为造成影响。

2.3 链接预测

传统的链接预测方法主要分为基于节点属性相似性、基于网络结构相似性、基于节点相似性这3类,代表算法有Jaccard Coefficient(Jaccard)[17]、Admic/Adar(AA)[18]等,大多基于节点类型和链接类型唯一的同质信息网络,而实际世界的信息网络中存在不同类型的节点,并且节点之间也存在不同类型的关系,即异质信息网络。同时,大规模稀疏的信息网络对链接预测中社会网络数据计算提出了挑战,近年来,许多学者提出的网络表示学习方法可以将信息网络中的节点表示成低维、稠密的向量形式,从而保留丰富的网络信息,在向量空间中具有表示及推理的能力。

网络表示学习方法主要分为3类:①基于矩阵因式分解的图嵌入。用矩阵表示节点间的关系,对矩阵(如邻接矩阵、拉普拉斯矩阵、节点属性矩阵等)进行分解得到节点的嵌入向量,可以解决矩阵稀疏化问题,传统的算法有Laplacian Eigenmap[19]、Graph Factorization[20]等,近几年来还有多采用高阶数据邻近矩阵以保留图结构的HOPE[21]、GraRep[22]等算法。②基于随机游走的方法。随着Word2vec模型的提出,针对图结构数据展开了基于随机游走的方法,通过生成节点序列来学习节点表示形式,再对生成的节点序列进行嵌入,最早的基于随机游走的表示方法是B. Perozzi等[23]提出的DeepWalk算法,随后Node2vec[24]算法采用了灵活的偏差随机游走策略,同时考虑了广度优先搜索(BFS)和深度优先搜索(DFS)。Struc2vec[25]算法则是从空间结构相似性的角度来对节点相似度进行判断。③基于神经网络的图嵌入。这类算法多将神经网络和图结构数据结合起来,主要有LINE[26]、GCN[27]等算法。

综上所述,微博传播预测研究还存在以下不足:

①没有区分用户在微博传播中的不同参与行为;②目前的传播预测分析方法多适用于同质信息网络;③在针对公安领域的微博传播预测中,尚未考虑事件的关键字特征。鉴于 GATNE 模型[28]可用于处理异质信息网络,且适用于大规模数据建模,同时能包含丰富的节点属性和网络结构特征,与社交网络中存在的多类节点以及节点间存在多类关系较为相符。因此,本文针对这些不足提出基于特定属性异质信息网络嵌入模型,从而对公安微博的转发、评论、关注等传播活动进行链接预测,根据用户和微博、用户和用户之间的关系构建异质信息网络,提取的节点属性特征包含用户特征、文本结构特征、事件特征、案件关键字特征,最后结合传播网络结构和节点属性特征,引入 GATNE-I 模型进行微博传播预测。

3 研究方法

本文将公安微博传播网络看作一种异质信息网络,网络中的节点代表用户或微博条目,节点属性包含用户特征、微博文本结构特征、案件特征、时间特征等,针对公安微博这一特定领域的微博传播问题,引入异质信息网络嵌入模型 GATNE-I 构建了基于特定属性异质网络嵌入的微博传播预测模型,添加了用户、微博、时间特征作为特定属性嵌入到 GATNE-I 模型中。首先对通缉微博及转发、评论用户数据进行预处理,提取出用户特征、时间特征,并识别出通缉微博中的案件地点、时间、人名等案件关键字特征。利用 xgboost 算法[29]实现对微博特征和用户特征的重要性进行排序。同时,构建基于转发关系、评论关系、关注关系的异质信息网络,网络中节点类别分为用户和微博,边的类别为用户与微博之间的转发、评论关系及用户与用户之间的关注关系,结合节点的属性特征,构建基于特定属性异质网络嵌入预测模型,从而对微博传播进行预测。

3.1 微博特定属性异质网络的构建

3.1.1 微博异质信息网络

设一个有向网络图 $G=(V,E,A)$,其中节点类型映射函数 $\varphi:V\rightarrow O$,边类型映射函数 $\Phi:E\rightarrow R$,其中 O 和 R 分别代表所有节点类型和边类型的集合。每个节点 $v\in V$ 都属于一个特定的节点类型,$A=\{x_i|v_i\in V\}$ 是所有节点的节点特征集合,其中 x_i 是节点 v_i 的关联特征;$E=\cup_{r\in R}E_r$ 表示 E_r 包含所有边类型,$r\in R$ 且 $|R|>1$,对于每个边类型 $r\in R$,我们将网络分割为 $G_r=(V,E_r,A)$,并且称其为特定属性异质信息网络。

本研究将微博异质信息网络抽象为包含两种节点:微博(I)与用户(U),以用户转发、评论、关注 3 种关系作为元路径,构建微博异质信息网络,如下:①基于微博转发网络的元路径。当某条微博被两个不同的用户转发时,元路径表示为 U1-R1-I-R2-U2,其中 U1、U2 表示不同的用户转发了同一条微博 I,R1 和 R2 分别表示用户 U1、U2 转发微博 I 后的转发微博。②基于微博评论网络的元路径。当某条微博被两个不同的用户评论时,元路径表示为 U1-C1-I-C2-U2,其中 U1、U2 表示不同的用户评论了同一条微博 I,C1 和 C2 分别表示用户 U1、U2 评论微博 I 后的微博评论。③基于用户关注网络的元路径。当某个用户被两个不同用户同时关注时,元路径表示为 U1-U-U2,其中 U1、U2 表示不同的用户关注了同一个用户 U。

3.1.2 特征提取

本文构建的通缉微博信息的传播预测模型主要从用户特征、微博特征、时间特征 3 个维度探讨影响通缉微博信息传播的因素,如表 1 所示:

表 1 通缉微博特征及特征值

认证类型		无认证/个人认证/机构认证
用户特征	粉丝数量	0－999/100 0－999 9/10 000－99 999/100 000－999 999/1 000 000－1000万/1 000 万以上
	关注数量	0－299/300－599/600－899/900－1 999/2 000－9 999/10 000－20 000
	发布微博总数	0－999/1 000－9 999/10000－99 999/10 万以上
	所在行业	传统媒体/自媒体/新媒体/政府机构/公众人物/个人团体组织/企业/公益组织/所在行业－其他
	是否为公安系统	是/否
	性别	男/女
	等级	是否为 VIP
	所在地区	北京/上海/广东/湖南/浙江/湖北…
微博特征	文本结构	是否有 URL/哈希标签/图片/视频/提及/表情
	案件关键字	是否包含地点关键字
		是否包含时间关键字
		是否包含行为关键字
		是否包含嫌疑人描述关键字
		是否包含通缉令等级
		有无悬赏
		案情类别
		案情进展
时间特征	所在时间段	深夜(00:01－06:00),清晨(06:01－08:30),上午(08:31－12:00),中午(12:01－14:00),下午(14:01－18:00),晚上(18:01－24:00)
	节假日	节假日;非节假日
	星期	周一;周二;周三;周四;周五;周六;周日
	首发微博	首发微博;非首发微博

（1）用户特征。本文将用户特征划分为用户的基本属性和用户行为特征。用户的基本属性包含用户认证类型、所在行业、是否为公安系统用户、性别、等级、所在地区 6 个维度，其中用户认证类型的特征值包含机构认证用户、个人认证用户和无认证用户，机构认证用户本身具有一定的影响力和权威性，当其发布微博时更容易引起大众的关注和讨论；用户所在行业分为传统媒体、自媒体、新媒体、政府机构、公众人物、个人团体组织、企业、公益组织、所在行业－其他，其中个人团体组织指粉丝团、同城会、老乡会等组织，将认证信息中或者用户名称中含有"公安""警察"等字样的用户归为公安系统用户；等级分为是否开通了微博会员，所在地区通过微博用户资料中的地理信息获取，最终选择的特征值包括中国 34 个省级行政区，以及"海外""其他"共 36 个特征值。由于用户的粉丝数和关注数可能会影响微博流行度[30]，本文也选取了用户粉丝数、用户关注数和用户已发布的微博数这 3 个特征作为用户行为特征。

（2）微博特征。综合考虑通缉微博的特殊性，本文将微博特征分为微博文本结构特征和案件关键字，在微博文本结构特征中，URL、哈希标签与微博转发具有强相关性[11]，而且用户在发布微博时，通常会增加图片、视频来传递更多的信息，观察通缉微博内容发现，微博用户在发布重大通缉信息时，通常会加上表情来强调内容信息，比如[话筒][震惊]等，因此本研究将是否有链接、哈希标签、图片、视频、提及（@）、表情纳入微博文本结构特征中。

《公安机关办理刑事案件程序规定》第二百六十六条明确指出："通缉令中应当尽可能写明被通缉人的姓名、别名、曾用名、绰号、性别、年龄、民族、籍贯、出生地、户籍所在地、居住地、职业、身份证号码、衣着和体貌特征、口音、行为习惯，并附被通缉人近期照片，可以附指纹及其他物证的照片，除了必须保密的事项以外，还应当写明发案的时间、地点和简要案情"。现在一般采用照片和文字合一的方式通报犯罪嫌疑人信息。因此通常公安微博发布的通缉令所包含的嫌疑人信息量较多，但由于在通缉微博传播过程中，其他用户并不需要按照这一规定发布通缉信息，因此发布的微博中所包含的信息量较少，如缺少包含嫌疑人文字信息的照片等，而当一条微博包含信息量较多时，往往更能引起大众的关注和讨论。因此在微博内容特性中，本研究添加了案件案情关键字特征，包含发案的时间关键字、地点关键字、行为关键字、嫌疑人描述关键字、通缉令等级、有无悬赏、案情类别、案情进展，其中通缉令等级、有无悬赏等可以直接通过关键字筛选进行判断，而时间关键字、地点关键字和行为关键字等可以通过命名实体识别进行抽取，在此基础上，本研究依据《中华人民共和国刑法》上的罪名类别对每条微博中包含通缉案件的类型进行标注。由于案件进行到不同阶段，嫌疑人状态可能从刚开始的在逃到随后的被捕，因此根据通缉微博所展示的当前嫌疑人状态对案件进展也进行了区分，分为在逃、被捕、自首等。

（3）时间特征。微博发布时间的不同，能接收到信息的用户数量也不同，因此本研究在考虑微博发布时间段、星期、法定节假日的基础上，根据微博用户作息规律将微博发布所在时间段划分为深夜（00:01－6:00）、清晨（6:01－8:30）、上午（8:31－12:00）、中午（12:01－14:00）、下午（14:01－18:00）、晚上（18:01－24:00）6 个阶段[8]。由于本研究选取的数据时间跨度较大，并且包含多个通缉事件，通缉令一经公安部发布可能被其他用户转发，尤其是容易引起社会恐慌的案件，嫌疑人在逃对社会仍具有威胁，更容易被传统媒体、自媒体等转发报导，因此本研究将首发微博定义为在数据集中同一通缉案中不同阶段下（如被捕、自首等）最先发布通缉信息的微博。

3.2 特征重要性排序

特征重要性是通过对数据集中的每个属性进行计算并排序而得出，其原理是一次随机为数据集抽取数据的一个特征，计算其性能指标的下降程度，变化越大，则代表特征就越重要。多采用随机森林、决策树、xgboost 等集成学习算法，大致分为提升法和套袋法。xgboost 算法是在 gbdt 的基础上对提升算法进行改进，对数据残差进行拟合，并在损失函数上加入了模型复杂度的正则项，能有效防止过拟合，同时并行和分布式设计使得算法具有非常快的训练速度。xgboost 模型把缺失值当做稀疏矩阵来对待，本身在节点分裂时并不考虑缺失值的数值。缺失值数据会被分到左子树和右子树分别计算损失，选择结果较优的子树。如果训练中没有数据缺失，预测时出现了数据缺失，那么默认被分类到右子树。

本研究采用 xgboost 算法进行特征重要性排序，采用 python 中 sklearn 包中的 train_test_split( ) 函数随机划分训练集和测试集，在此将链接预测问题看成一个二分类问题，两个节点之间相连则设置标签为 1，否则设置为 0。xgboost 算法通过不断添加 CART 树来学习一个新函数，拟合上次预测的残差。xgboost 的输入是

孙冉，安璐.异质信息网络嵌入视角下公安微博传播预测研究[J].图书情报工作,2020,64(21):67－76.

两个节点的特征向量的组合，比如转发预测中，拟输入转发用户的特征值和微博的特征值。对于多维特征向量 x_i，则 xgboost 的输出如公式(1)所示：

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), f_k \in F \quad 式(1)$$

其中，k 是 CART 树的棵数，F 表示所有可能的 CART 树，$f_k(x_i)$ 表示 CART 树 k 的分类结果。xgboost 模型的目标函数如公式(2)所示：

$$obj(\theta) = \sum_i^n l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \quad 式(2)$$

其中，目标函数的第一项 l 为损失函数，度量预测值与目标值之间的差，第二项 Ω 为代表正则项，为 k 棵 CART 树的复杂度之和，包括叶子结点的个数和叶子结点的分数。

3.3 GATNE－I 模型

Y. Cen 等[28]于 2019 年提出的 GATNE 模型(GATNE-I 和 GATNE-T)能够用来处理真实世界中由大规模节点和多种类型的边组成的网络，而且网络中的每个节点都能与不同的属性相关联。本文引入 GATNE-I 模型来解决异质信息网络中的不同节点不同边的链接预测问题，对公安微博的转发、评论、关注等传播活动进行链接预测。模型中每个节点在每个边类型下的嵌入由基于节点特征的通用嵌入 b_i、边嵌入 u_i 和节点属性构成，分别对应描述结构信息、异质信息和属性信息。节点的通用嵌入是每个节点在每种边类型下共享的，而每个节点的边嵌入则是按照不同的边类型通过相邻节点的边嵌入计算得到的，节点 v_i 在边类型为 r 下第 k 层的边嵌入计算方式如公式(3)所示：

$$u_{i,r}^{(k)} = aggregator(\{u_{j,r}^{(k-1)}, \ \forall v_j \in N_{i,r}\}) \quad 式(3)$$

其中 i,j 表示异质信息网络中的节点编号，r 表示边类型，聚合函数可以采用平均聚合或者其他类型的池化聚合。在模型 GATNE-T 中，每个节点 v_i 在每种边类型 r 下初始边嵌入 $u_{i,r}^{(0)}$ 是随机初始化的。将第 k 层的边嵌入 $u_{i,r}^{(k)}$ 定义为 $u_{i,r}$ 并将节点 v_i 的所有边嵌入连接成大小为 $s*m$ 的矩阵 U_i，其中 s 表示边嵌入的维度，m 表示边类型的数量。

通过自注意机制来计算 U_i 中在每种边类型 r 下向量之间线性组合的系数 $a_{i,r}$，从而得到每种边类型下的向量表示对各个边类型的权重，如公式(4)所示：

$$a_{i,r} = softmax(w_r^T tanh(W_r U_i))^T \quad 式(4)$$

其中 w_r 和 W_r 分别是大小为 d_a 和 $d_a * s$(d 为通用嵌入的维度)下边类型 r 的可训练参数，T 表示向量或矩阵的倒置。最终得到 GATNE-T 模型中，每个节点 v_i 在边类型 r 下的向量表示，如公式(5)所示。其中 b_i 是节点 v_i 的通用嵌入，α_r 是一个超参数，用来表示边嵌入相对于整体嵌入的重要性，$M_r \in R^{s*d}$ 是一个可训练的变化矩阵。

$$v_{i,r} = b_i + \alpha_r M_r^T U_i a_{i,r} \quad 式(5)$$

GATNE-T 模型不能处理没有出现的节点，而在实际情况中的网络数据很多是不全面的，而引入了节点特征的 GATNE-I 模型可以解决这一问题，将通用嵌入 b_i 定义为节点 v_i 的属性 x_i 的参数方程。不同节点 v_i 的属性 x_i 可能具有不同维度。原先在 GATNE-T 模型中随机初始化的 $u_{i,r}^{(0)}$ 则是通过节点的属性函数得到，如公式(6)所示，其中 $g_{z,r}$ 也是一个变换函数，用来将特征变换到节点 v_i 在边类型 r 的边嵌入。

$$u_{i,r}^{(0)} = g_{z,r}(x_i) \quad 式(6)$$

同时在 GATNE-I 模型中会在节点 v_i 在边类型 r 上的整体嵌入中增加一个额外的属性项，节点 v_i 在某种边类型 r 下的向量表示 $v_{i,r}$ 的表达如公式(7)所示，其中 β_r 是系数，D_z 是节点 v_i 对应节点类型 z 上的特征转换矩阵。

$$v_{i,r} = h_z(x_i) + \alpha_r M_r^T U_i a_{i,r} + \beta_r D_Z^T x_i \quad 式(7)$$

4 实验及结果分析

4.1 数据集

本文采用的数据集来自国内社交媒体平台新浪微博，以检索词为“通缉 嫌疑”采集原创微博及其评论、转发和所有用户的基本信息，并且经过人工校验去除与通缉微博无关的微博，共采集了 2016 年 1 月 1 日－2019 年 9 月 15 日的 14 905 条原始微博、86 146 条转发微博、62 548 条微博评论。选取发布原始微博的用户，采集其关注与被关注用户信息，由于数据量过大，并且关注网络过于稀疏，去除在关注网络中只出现过一次的用户，得到 16 8059 条关注与被关注信息，共包含了 14 3370 个微博用户。在微博异质信息网络中，节点代表微博用户，边代表用户之间的转发、评论、关注等关系。采集的微博用户字段包含用户 id、用户名、性别、省份、VIP、认证、发文数、关注数和粉丝数；微博文本信息字段包含微博 id、用户 id、发布时间、登录设备、点赞数、转发数、评论数、图片链接和微博文本。

4.2 通缉信息的传播模式分析

以通缉微博转发网络为例，本文对信息传播中不同主体发布的微博中用户属性进行探讨，由于微博转发网络结构较大，本研究选择数据集中出现一次以上的节点，得到 697 个节点、5 118 条边的网络结构数据。其中，共计 108 个公安系统用户发布的微博中有 323 个公安系统用户参与转发。经过 gephi 软件可视化发

现，同一省（直辖市）内，不同地区的公安用户通常会转发地区级别最高的公安用户发布的微博，转发网络中的普通用户也多为同一省（直辖市）内的微博用户。以“平安重庆”发布的微博为例，如图 1 所示，发现其转发用户多为“平安南岸”“平安黔江”等重庆直辖市下不同地区的公安用户，普通用户如“优雅的猫 forever”“雾都老猫”也属于重庆地区。这说明通缉令信息的传播范围一般局限在省市内（直辖市），不同省市（直辖市）之间公安微博对通缉信息的传播较少。



图 1 “平安重庆”的转发网络

以“人民日报”为例，相比于公安系统用户，转发网络中普通用户较多，还有部分公安系统用户，如“武昌白沙洲派出所”“平安吉利”等，公安系统用户的所在地区较为分散，这也侧面表明新闻媒体因其本身具有一定的影响力，通缉信息的传播突破了以公安微博为中心的用户群体。

4.3 案件关键字提取

通缉微博文本中包含大量与通缉事件相关的信息，包括案件发生的时间、地点、人物（姓名）、犯罪行为等，对于有些案件关键字（如通缉令等级），可以直接对微博文本进行关键词筛选来判断。由于现有的通缉令多以照片文字二合一的方式进行传播，本文根据爬取的微博图片链接下载图片，再利用闪电 OCR 图片文字识别软件得到图片中的文本信息，结合微博文本内容，采用 python 中的 hanlp 分词工具对原创微博文本进行分词，并进行词性标注，识别出人名、地名、时间等信息，如覃志钢/n、广东/ns、23 日/t 等。对属性中案件关键词字段进行补充如表 2 所示，案情类别是依据《中国人民共和国刑法》中的罪名及定义，结合案件嫌疑人行为关键字等进行划分。

4.4 微博特征重要性排序

为了判断不同节点的网络关系中特征值对链接预测的影响，本研究采用 xgboost 算法筛选出重要特征，分别对转发预测和评论预测中的特征重要性进行排序，结果见图 2。

实验结果显示，在转发预测和评论预测中，微博案件关键字特征的重要性均明显高于其他特征的重要性，表明含有通缉案件相关信息的微博更容易被转发或评论，公安微博或者新闻媒体发布通缉信息时，对案件补充的信息越多，越容易引发用户参与信息传播；用户粉丝数、节假日的重要性均明显低于其他特征的重要性，这说明用户的粉丝数和是否为节假日对预测用户转发或评论通缉微博的影响较小；用户所在地、星期、用户所在行业、微博发布时间段的特征重要性都较高（大于 0.05）。但用户认证类型的重要性在转发预测中相对于在评论预测中要低，首发微博的重要性在转发预测中相对于在评论预测中则高许多，这也反映出在同一个通缉事件中发布微博的时间越早就越容易被用户转发。

表 2 案件关键词字段

案件关键字	字段
地点关键字	广西/ns、环江县/ns、大谭镇/ns、赵屯村/ns、北京站/ns 等
时间关键字	2015 年/t、10 月/t、23 日/t、日前/t、近日/t、凌晨/t、当晚/t 等
行为关键字	杀害/v、抢劫/v、涉黑/v、殴打/v、放贷/v、奸杀/v 等
嫌疑人描述关键字	皮肤/n、圆脸/n、秃顶/n、肤色/n、体态/n、牛仔裤/n 等
通缉令等级	A 级通缉令、B 级通缉令、国际通缉令
有无悬赏	悬赏/v、奖励/v、奖赏/vn、奖征/n 等
案情类别	组织、强迫、引诱、容留、介绍卖淫罪；走私罪；走私、贩卖、运输、制造毒品罪；制作、贩卖、传播淫秽物品罪；危害税收征管罪；危害公共卫生罪；危害公共安全罪；贪污贿赂罪；生产、销售伪劣商品罪；扰乱市场秩序罪；扰乱公共秩序罪；侵犯公民人身权利、民主权利罪（故意杀人罪、故意伤害罪、强奸罪等）；侵犯财产罪（抢劫罪、盗窃罪等）；破坏金融管理秩序罪；破坏环境资源保护罪；金融诈骗罪；妨害文物管理罪；妨害司法罪；妨害国（边）境管理罪；妨害公务罪；妨害对公司、企业的管理秩序罪；渎职罪
案情进展	投案自首/nz、在逃/nz、抓捕归案/n、被捕/v 等

从微博的角度来看，以微博案件关键字和时间特征、文本结构特征中部分特征进行具体分析，如图 3 所示。在案件关键字特征重要性排序中，A 级通缉令（0.033）、破坏金融管理秩序罪（0.029）、案件类别－其他（0.055）对微博评论预测的影响较大，而地点关键词（0.027）、嫌疑人描述关键词（0.022）、侵犯财产罪

孙冉，安璐. 异质信息网络嵌入视角下公安微博传播预测研究[J]. 图书情报工作,2020,64(21):67－76.

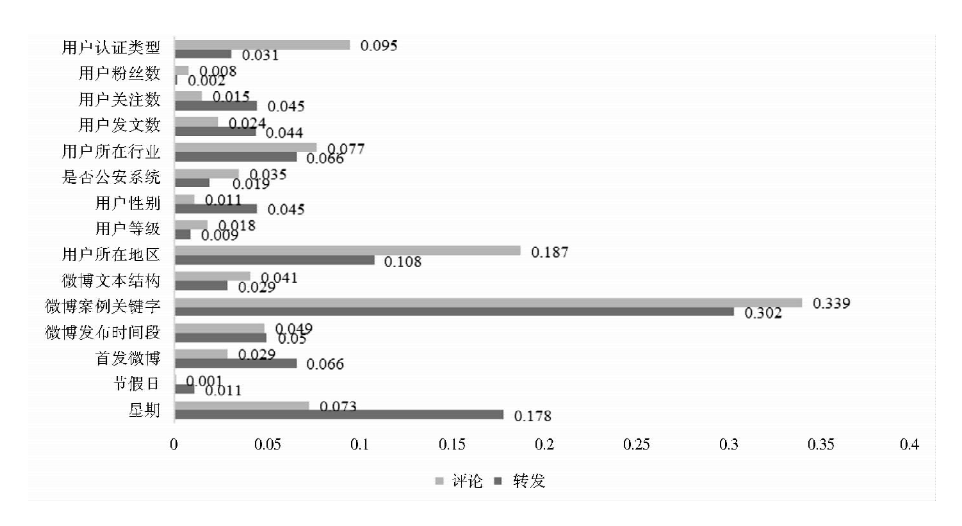


图 2　特征重要性排序

(0.028)、被捕(0.028)等对微博转发预测的影响较大。在微博时间特征中,周日(0.138)在转发预测中的重要性最高,在评论(0.017)预测中的重要性也较高,首发微博在转发和评论预测中的重要性都较高(大于0.02),晚上、周二、周六在转发预测中的重要性较低于在评论预测中。在微博文本结构特征中,视频和哈希标签的特征重要性均较高,这给公安系统用户发布通缉微博时一些启示:发布的微博尽可能应该包含案件相关信息,如嫌疑人的特征、案件进展等,发布微博的时间可以选在周六、周日,发布的微博通过添加与案件相关的视频、图片、话题标签等来传递更丰富的信息,从而引发更多用户的关注,如公安部在2019年7月采取在全国范围通缉50名重大在逃人员的行动中,在逃人员来自全国各地,当各地公安发布微博时,会添加话题#公安部通缉50名重大在逃人员#,虽然本地公安微博自身影响力较低,但也能吸引较多的转发。

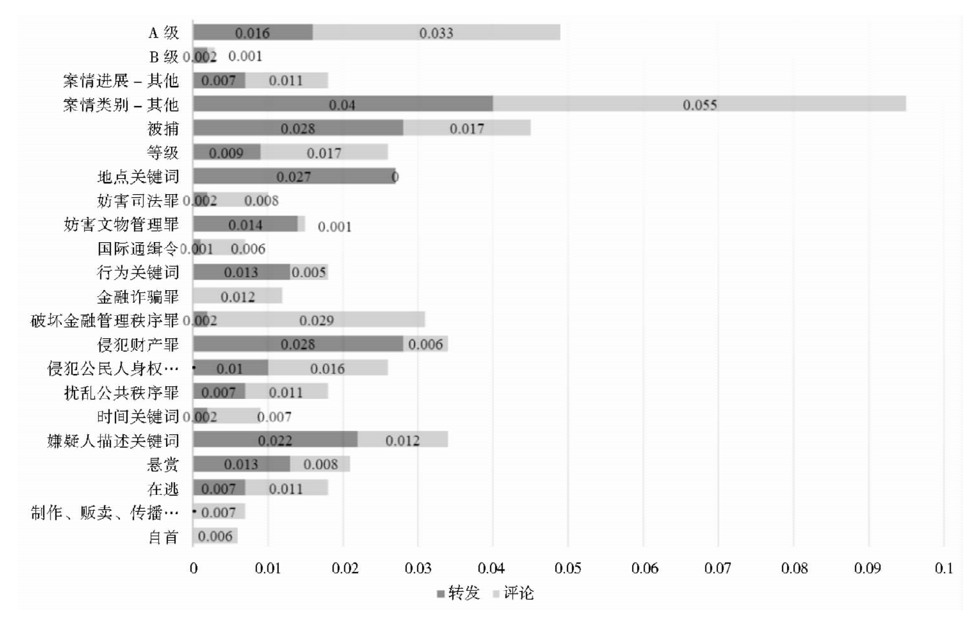


图 3　微博－案件关键字特征重要性

从参与转发或者评论的用户角度来看,用户所在地区在转发(0.180)和评论(1.87)预测中的重要性都最高,其中,天津、北京、河北等华北地区在转发和评论预测中的重要性均较高,而广西、海南、香港和澳门等华南地区在转发和评论预测中的重要性均较低。机构认证、政府机构、公安系统特征对转发、评论预测的影响较大,也侧面反映出通缉微博传播的用户圈仍然具有一定的局限性,政务微博之间的合作传播较为明显。关注数为300－599和关注数为2 000－9 999的用户特征在转发预测中的重要性明显高于在评论预测中的重要性,但是关注数为900－1 999的用户特征对转发预测的影响相比对评论预测的影响要低。这有助于新

浪等社交媒体平台进行用户推荐,针对容易转发、评论通缉微博的用户群体推送相关微博,既能增加公安用户等政务微博的影响力,促使通缉信息能得到广泛传播,也帮助广发用户贡献自己的力量协助公安部门惩恶扬善。

4.5 实验结果

本研究基于特定属性异质信息网络嵌入(GATNE)模型进行链接预测实验,为了评价模型的性能,本文选择相同的实验数据,分别构建 DeepWalk、Node2vec、Line、GAE、SDNE 等链接预测模型,并将这些模型的性能评估结果与 GATNE-I 模型进行比较,由于 DeepWalk、Node2vec 等模型不能对不同边类型的网络进行处理,因此本文对转发、评论、关注关系的网络分别采用基线模型进行预测。本研究先抽取部分数据集进行实验,如表 3 所示。数据集中正负样本的比例会影响预测的准确率,一般将原始数据集划分为训练集(70%)、测试集(约 20%)和验证集(约 10%),并且每个集合中的正负样本数量应大致相等。实验环境为 2 * Intel(R) Xeon(R) E5 - 2640 v4 x86_64,2.4GHz,20 核心,Nvidia Tesla V100,内存 16G。各模型的性能评估结果如表 4 所示。其中,AUC 是指 ROC 曲线下的面积,ROC 曲线是以假正例率(FPR)和真正例率(TPR)作为变量而做出的曲线,其中 FPR 为横坐标,TPR 为纵坐标;F1 是模型精确率和召回率的调和平均数,如公式(8)所示;PR 曲线是以精确率(precision)和召回率(recall)作为变量而做出的曲线,其中 recall 为横坐标,precision 为纵坐标,则 PR 值是指 PR 曲线下的面积;AUC、F1、PR 数值越大,则代表模型性能越好。

$$F1 = 2 * \frac{precision * recall}{precision + recall} \quad 式(8)$$

表 3 抽取的部分数据集

数据集	节点数(个)	边数(条)
转发	3 534	5 127
评论	2 576	2 699
关注	5 472	81 546

表 4 模型的性能评估结果(部分数据集)

模型	转发指标			评论指标			关注指标		
	AUC	F1	PR	AUC	F1	PR	AUC	F1	PR
SVD	0.552	0.524	0.536	0.672	0.625	0.618	0.736	0.675	0.724
DeepWalk	0.562	0.547	0.543	0.693	0.664	0.618	0.766	0.680	**0.782**
Node2vec	0.562	0.562	0.542	0.675	0.647	0.599	0.736	0.652	0.757
Graph Factorization	0.559	0.550	0.550	0.601	0.583	0.574	0.719	0.662	0.733
LINE	0.660	0.599	0.687	0.631	0.587	0.633	**0.770**	**0.693**	0.780
GAE	0.566	0.587	0.552	0.622	0.606	0.600	0.604	0.634	0.540
SDNE	0.704	0.608	0.680	**0.760**	0.669	0.770	0.737	0.678	0.728
GATNE-T	**0.712**	0.646	0.739	0.711	0.610	0.752	0.638	0.599	0.609
GATNE-I	0. 683	**0.688**	**0.764**	0.732	**0.675**	**0.783**	0.618	0.594	0.578

从表 4 看出,对比于基线模型,在转发、评论关系的数据集上 GANTE-T 模型具有较高的准确率,同时,由于节点的特征非常丰富,因此 GATNE-I 的效果明显好于 GATNE-T。虽然在关注网络上的预测效果较差,但是整体来看,GATNE-I(T)模型效果较好。随后在全数据集上进行实验,并且选择 DeepWalk、Node2vec 作为基线模型进行对比,得到实验结果如表 5 所示,设置 GATNE-I 模型的词向量维度为 200,随机游走序列长度为 10,每个节点选择 20 个随机游走序列,窗口大小为 5,负样本数为 5。

表 5 模型的性能评估结果(全数据集)

模型	转发指标			评论指标			关注指标		
	AUC	F1	PR	AUC	F1	PR	AUC	F1	PR
DeepWalk	0.456	0.468	0.486	0.456	0.424	0.594	0.733	0.668	0.756
Node2vec	0.528	0.520	0.551	0.442	0.404	0.589	**0.744**	**0.677**	**0.768**
GATNE-T	0.685	0.639	0.662	0.716	0.644	0.737	0.603	0.574	0.571
GATNE-I	**0.737**	**0.651**	**0.672**	**0.799**	**0.689**	**0.792**	0.569	0.543	0.577

上述实验结果中,GATNE-I 模型在评论、转发预测数据集上的 AUC 分别达到 0.799、0.734,同样的,GATNE-I(T)模型在关注数据集上的表现较差,但是整体来看,GATNE-I(T)模型的模型效果仍优于 DeepWalk、

Node2vec 模型。DeepWalk、Node2vec 模型是基于同质信息网络,而 GATNE-I(T)模型能对异质信息网络进行处理,并且能有效处理大规模数据,更加符合现实世界中数据规模庞大的社交网络。

5 结论

本文针对通缉微博这一特定领域信息传播问题,从用户属性特征、微博案件关键字特征、微博文本结构特征、时间特征等方面,基于微博和用户视角,采用 xgboost 算法评估特征对微博传播预测的重要性,包括用户的转发、评论,研究发现微博案件关键字特征不管是在转发行为预测还是评论行为中的重要性都是最高。随后,本文探讨了公安用户和媒体用户的转发网络中用户的基本属性,发现公安用户发布的通缉微博的传播圈较为局限,省市(直辖市)内各地区的公安微博之间合作较紧密。另一方面,本文构建了基于转发、评论、关注关系的异质信息网络,并且根据提取的用户和微博属性,提出了基于特定属性的异质网络嵌入模型,预测通缉微博的转发、评论以及用户之间的关注关系。实验结果表明,模型在转发、评论预测中准确度分别达到了 0.734 和 0.799,高于其他基线模型,有助于公安机关更好地践行“网络通缉”的举措,更好地通过用户个性化推荐来促进网民的积极参与,为公安机关最大化地利用社交网络提供现实条件。本研究的不足之处在于:①本文将多级转发、评论数据均看成一级转发、评论,未来可以更深入探索多级转发、评论中不同用户之间的关系;②本文提出的预测模型是基于异质信息网络,但现实世界中的信息传播过程是动态变化的,没有将时间因素考虑在内;③在对不同主体发布的微博中传播用户的属性进行探讨时,没有考虑到不同等级的通缉案件会引发省际间不同的传播模式,未来可进一步展开研究。

致谢:感谢图书情报国家级实验教学示范中心为本研究提供的实验支持!

参考文献:

[1] 郑建国,朱君璇,曹如中. 基于情境的社交网络信息传播链路预测研究[J]. 情报理论与实践,2018,41(6):94－99.
[2] 肖飞. 公共危机事件中政务微博的舆情信息工作理念与策略探析——以雅安地震为例[J]. 图书情报工作,2014,58(1):44－47,71.
[3] 刘小平,田晓颖. 传统媒体与新媒体微博社会网络特征对比分析实证研究[J]. 图书情报工作, 2018, 62(5): 106－114.
[4] 张连峰,周红磊,王丹,等. 基于超网络理论的微博舆情关键节点挖掘[J]. 情报学报, 2019, 38(12): 1286－1296.
[5] 陈然,刘洋. 基于转发行为的政务微博信息传播模式研究[J]. 电子政务, 2017 (7): 108－117.
[6] LI L, ZHANG Q, TIAN J, et al. Characterizing information propagation patterns in emergencies: a case study with Yiliang Earthquake[J]. International journal of information management, 2018, 38(1):34－41.
[7] 陈贵梧. 地方政府创新过程中正式与非正式政治耦合研究——以公安微博为例[J]. 公共管理学报,2014,11(2):60－69,141－142.
[8] 安璐,易兴悦,孙冉. 恐怖事件情境下微博影响力的预测及演化[J]. 图书情报知识,2019(4):52－61,81.
[9] 徐月梅,刘韫文,蔡连侨. 基于深度融合特征的政务微博转发规模预测模型[J]. 数据分析与知识发现, 2020, 4(2/3): 18－28.
[10] BOYD D M, ELLISON N B. Social network sites: definition, history, and scholarship[J]. Journal of computer - mediated communication, 2007, 13(1): 210－230.
[11] 黄微,刘熠,许烨婧,等. 网络舆情推文的热度测度模型构建[J]. 图书情报工作,2019,63(20):17－25.
[12] SUH B, HONG L, PIROLLI P, et al. Want to be retweeted? large scale analytics on factors impacting retweet in twitter network [C]//Proceedings of the 2010 IEEE second international conference on social computing. Washington, DC: IEEE Computer Society, 2010: 177－184.
[13] 田磊,任国恒,王伟. 面向阅读推广的微博用户转发行为为预测[J]. 情报学报,2017,36(11):1175－1182.
[14] ZHU J, XIONG F, PIAO D, et al. Statistically modeling the effectiveness of disaster information in social media[C]//Proceedings of the 2011 IEEE global humanitarian technology conference. Seattle: IEEE Computer Society, 2011: 431－436.
[15] JIANG B, LIANG J, SHA Y, et al. Retweeting behavior prediction based on one-class collaborative filtering in social networks[C]//Proceedings of the 39th international ACM SIGIR conference on research and development in information retrieval. New York: Association for Computing Machinery, 2016: 977－980.
[16] KHAN M L. Social media engagement: what motivates user participation and consumption on YouTube? [J]. Computers in human behavior, 2017, 66: 236－247.
[17] JACCARD P. The distribution of the flora in the alpine zone. 1 [J]. New phytologist, 1912, 11(2): 37－50.
[18] ADAMIC L, ADAR E. How to search a social network[J]. Social networks, 2005, 27(3): 187－203.
[19] BELKIN M, NIYOGI P. Laplacian eigenmaps and spectral techniques for embedding and clustering[C]//Proceedings of the 14th international conference on neural information processing systems: natural and synthetic. Cambridge: MIT Press, 2002: 585－591.
[20] AHMED A, SHERVASHIDZE N, NARAVANAMURTHY S, et al. Distributed large-scale natural graph factorization[C]//Proceedings of the 22nd international conference on World Wide Web. New York: Association for Computing Machinery, 2013: 37－48.

[21] ZHU D, CUI P, ZHANG Z, et al. High-order proximity preserved embedding for dynamic networks[J]. IEEE transactions on knowledge and data engineering, 2018, 30(11): 2134-2144.

[22] CAO S, LU W, XU Q. Grarep: learning graph representations with global structural information[C]//Proceedings of the 24th ACM international on conference on information and knowledge management. New York: Association for Computing Machinery, 2015: 891-900.

[23] PEROZZI B, Al-RFOU R, SKIENA S. Deepwalk: online learning of social representations[C]// Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining. New York: Association for Computing Machinery, 2014: 701-710.

[24] GROVER A, LESKOVEC J. Node2vec: scalable feature learning for networks[C]//Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. New York: Association for Computing Machinery, 2016: 855-864.

[25] RIBEIRO L F R, SAVERESE P H P, FIGUEIREDO D R. Struc2vec: learning node representations from structural identity[C]//Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining. New York: Association for Computing Machinery, 2017: 385-394.

[26] TANG J, QU M, WANG M, et al. Line: large-scale information network embedding[C]// Proceedings of the 24th international conference on World Wide Web. Geneva: International World Wide Web Conferences Steering Committee, 2015: 1067-1077.

[27] KIPF T N, WELLING M. Semi-supervised classification with graph convolutional networks[C]// Proceedings of the 5th international conference on learning representations. 2017: 1-14.

[28] CEN Y, ZOU X, ZHANG J, et al. Representation learning for attributed multiplex heterogeneous network[C]//Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery and data mining. New York: Association for Computing Machinery, 2019: 1358-1368.

[29] CHEN T, GUESTRIN C. Xgboost: a scalable tree boosting system [C]//Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. New York: Association for Computing Machinery, 2016: 785-794.

[30] SUH B, HONG L, PIROLLI P, et al. Want to be retweeted? large scale analytics on factors impacting retweet in Twitter network [C]//Proceedings of the 2010 IEEE second international conference on social computing. Washington, DC: IEEE Computer Society, 2010: 177-184.

作者贡献说明:
孙冉:文献调研,实验与数据分析,论文撰写;
安璐:研究框架确定,论文撰写与修改。

Propagation Prediction of Police Microblog Entries Based on Heterogeneous Information Network

Sun Ran[1] An Lu[1,2]
[1] School of Information Management, Wuhan University, Wuhan 430072
[2] Center for Studies of Information Resources, Wuhan University, Wuhan 430072

Abstract: [**Purpose/significance**] This study aimed to predict whether microblog users would retweet or comment on the microblog entries containing wanted information. We also evaluated the important features that affected the spread of wanted microblog entries to help the public security departments improve their operation performance and enhance the communication and cooperation between the police and the public. [**Method/process**] Based on the characteristics of the wanted microblogging, we combined user features, time features and structure features, and extracted event features in microblog entries, such as location keywords, time keywords, the wanted level and so on. The Xgboost algorithm was used to calculate the importance of different features in the retweet and comment prediction. In combination with the features of transmission network and node attributes, we trained and evaluated a prediction model based on heterogeneous information network embedding. [**Result/conclusion**] The values of the AUC in retweeting and commenting data sets are 0.737 and 0.799 respectively. As the model integrated network structure characteristics and different nodes' attributes, it was closer to the heterogeneous information network in reality and had higher accuracy than the traditional link prediction model. In addition, the result of features' importance showed that the keyword features of the proposed event features had the highest importance among all the features that affected the prediction of microblog entries retweeted and commented.

Keywords: information dissemination public security microblog link prediction graph representation learning heterogeneous information network